DOCUMENT RESUME

ED 420 691 TM 028 366

AUTHOR Johanson, George; Alsmadi, Abdalla
TITLE Differential Person Functioning.
PUB DATE 1998-04-00
NOTE 9p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Counseling; *Diagnostic Tests; *Item Bias; Matrices; *Rating Scales; Test Items

ABSTRACT

In many testing situations, differential item functioning (DIF) is a potentially serious problem. It occurs when a test item appears to be easier for one group of examinees than another even after controlling for overall skill level. Differential person functioning (DPF) can occur when "items" can be considered raters and the persons are the objects being rated. This paper introduces the notion of DIF with object-rater data and transposed person-item matrices. The discussion only considers methods for DIF or DPF detection using a binary coding of raters and right-wrong coding of subjects on cognitive tests. Three examples are used: (1) data from a standard setting session of a test of counseling knowledge and skills using an Angoff variation with 9 expert judges or raters; (2) ratings by 14 committee members of 25 research proposals; and (3) data from a mathematics achievement test taken by 384 sixth graders. These examples serve to illustrate the potential utility of DPF, or the application of DIF to rating data and transposed data from cognitive or diagnostic assessment. Notions incorporated in DPF might be especially useful in the area of diagnostic testing. (Contains 4 figures and 17 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Differential Person Functioning

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
George Johanson
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

George Johanson and Abdalla Alsmadi
College of Education
Ohio University

A Paper Presented at the 1998 Annual Meeting of the
American Educational Research Association, San Diego, CA

ED 420 691

In many testing situations, differential item functioning (DIF) is a potentially serious problem. DIF occurs when a test item appears to be easier for one group of examinees than for another group even after controlling for an overall skill level (Dorans & Holland, 1993). The total score on the test is often used to form subgroups of individuals with similar overall scores or skill levels. That is, examinees from two groups (frequently referred to as the *focus* and *reference* groups) will both belong to a number of equal-ability subgroups; evidence of DIF is seen when the item in question is easier (or more difficult) for the focal group members than the reference group members within the equal-ability subgroups. DIF is said to be *uniform* or *consistent* (Camilli & Shepard, 1994) when the difficulty direction is the same, say, more difficult for the focal group members, across all of the equal-ability subgroups .

On the other hand, an item is said to have *impact* (Dorans, 1989) if it is simply more or less difficult for one group than the other. Impact does not require conditioning on an overall score and is not the same as DIF. Impact might simply reflect genuine focal-reference group differences. In fact, it might be quite unusual for a test item to not show impact in certain situations. For example, an item reflecting a fifth grade mathematics curriculum would be expected to show impact between a reference group of fifth graders and a focus group of fourth graders. The item would only be considered to function differentially, however, if fourth and fifth graders *with the same overall score on the entire mathematics test* perform differently on this item.

Item *bias* is also not the same as DIF. Camilli and Shepard (1994) state that "Ideally, DIF statistics would be used to identify all items that function differently for different groups; then, after logical analysis as to *why* the items seem to be relatively more difficult, a subset of DIF items would be identified as 'biased' and presumably eliminated from the test." (p. 16).

When objects (physical objects, persons, test items, performances, responses, or otherwise) are rated by a group of judges, the data are comparable to the usual person-item data matrix from a cognitive assessment. The 'items' can be considered raters and the 'persons' the objects being rated. Such object-rater data occur often in practice. Specific instances would be athletic competitions such as gymnastics (the gymnasts are the objects being evaluated and the judges are the raters); performance assessments (the object is the student performance and the teachers are the raters); standard setting (Crocker & Algina, 1986) using one of the popular Angoff modifications (the objects are the test items whose difficulties are being estimated and the judges are the raters); student evaluations of faculty (faculty are the objects and students are the raters). Since the rater or judge is more generally a person, this type of differential functioning might best be referred to as *differential person functioning* or, *DPF*.

Much of the current literature involving rater bias (we have not noted the phrase 'differential functioning' used with raters or persons) would actually seem to involve only rater impact. That is, there would seem to be little or no effort to condition the relationship between rater and object-group

TM028366

(focal or reference) by an overall rater measure, but more interest in impact or mean rater differences between object-groups (e.g., Ansorge & Scheer, 1988). In some cases, external influences on the ratings (such as the various forms social desirability) are described as causing bias (e.g., Keller & Bishop, 1985). This lack of clarity with respect to the
term 'bias', which is also commonly used in parameter estimation of all sorts, is an indication of the need for a more precise concept such as differential person functioning.

Unconditioned rater differences are often put forth as evidence of bias when these discrepancies might be better understood using the more precise language from cognitive assessment. For example, we might suppose that a judge in a figure skating competition evaluates most of the skaters from a particular country less favorably than those from other countries. This may not be bias or DPF as defined previously, but rather *impact*, a possibly reasonable measure of actual skill differences between the groups of skaters. If all skaters were put into subgroups of similar overall skating skills (using, say, a grouping based on the mean ratings of all of the judges) and in these classes of reasonably comparable skaters this judge favored members of one country over all others, *then* we might well consider this judge differentially functioning and, for many purposes, biased. The meaning of the term *bias* would be substantially different in these two contexts.

A somewhat different application of DPF would be in the area of achievement testing. That is, transpose the usual person-item data matrix to an item-person matrix; analyze this matrix as you might for DIF where the reference and focus groups are now item clusters representing, say, different content domains or item formats. Persons are investigated individually for differential functioning over these focal and reference item groups. If a person simply has better performance in one content area than another, this could be a form of person impact, a not unusual diagnostic outcome. If a person functioned differentially, however, this would indicate that there were content area performance differences, *after conditioning on, or controlling for,*
overall item difficulty. In those examples with transposed data matrices, DPF would seem to be the intuitive designation and would provide a new and possibly quite informative type of diagnostic information.

The purpose of the present paper is to introduce the notion of differential item functioning with object-rater data and transposed person-item matrices. We present examples both with and without differential functioning.

Method

Our discussion will consider only methods for DIF or DPF detection using a binary coding of raters (we will use above-average/below-average recoding where necessary) and right-wrong coding of subjects on cognitive tests. Of the empirical methods for identifying DIF in binary items, the Mantel-Haenszel (MH) procedure (Mantel & Haenszel, 1959; Dorans, 1989) is often recommended (Holland & Thayer, 1988; Dorans & Holland, 1993). When using the MH procedure with object-rater data, an overall rating (or score, including the item in question) is used to form a number (K) of subgroups of objects with similar total ratings. K is often just the number of all possible raw scores or ratings. For each level of K, a 2x2 frequency table if formed by crossing the judge's binary rating (above-average/below-average) with group membership (focus-reference). An overall odds-ratio is then computed from the comprehensive Kx2x2 table as a test statistic (approximately distributed as a χ^2 with one degree of freedom) testing the null hypothesis that the odds-ratio equals one. This null is the hypothesis of no DPF. Using the above-average/below-average interpretation of a judge's binary rating, the MH odds-ratio reflects the odds that an object in the focal group will be rated above-average by the rater or person under investigation when compared to the binary rating by the same person of an object in the reference

group (or vice-versa) when the objects in the reference and focal groups are reasonably similar (within subgroups) on overall ratings. More detailed explanations and formulae can be found in Raju, Bode, & Larsen (1989) or Camilli & Shepard (1994).

A number of standard statistical packages compute MH statistics. A macro for SPSS was recently made available (Nichols, 1994) and was used in this study. Uttaro and Millsap (1994) indicate that both the odds-ratio and the significance test are important and should be consulted in the detection of DIF.

Results

Example One

Standard setting or the setting of passing scores on achievement, certification, or other cognitive tests is now common practice. G. Cizek (1996) states that of the various methods used, "The Angoff method has become the most rigorously researched and widely used of the item-based procedures." (p. 20). This approach is due to W. H. Angoff (1984) and, with the usual modifications, requires a group of judges to estimate the likelihood or probability that a marginal candidate will succeed on each item of the test. The standard or passing score for the test can then be set to the mean of these probabilities. According to this method, a greater probability estimate for an item would indicate that a larger proportion of the marginal group would be expected to succeed on the item. That is, this item would be seen by the rater as a relatively easy item.

The data for this example were from a standard setting session of a test of counseling knowledge and skills using an Angoff variation (on a test with 99 items) and involving nine expert judges or raters. Each rater estimated the probability of success of a marginal candidate for each item and the mean of these (0.69) was the standard. The MH procedure requires binary variables from each rater. To accommodate this, the responses to each item were recoded 0-1 where an estimate greater than or equal to the passing score (0.69) was recoded '1' and a response of less than 0.69 was recoded '0'. A total binary score was then computed and used to group the 99 items or objects into (10) similar item difficulty subgroups.

The question of interest was whether any of the judges was performing differentially across different content areas. In particular, the 12 items corresponding to the content area of *Human Growth and Development* (the focal group of items) were contrasted with the remaining 87 items (the reference group of items). A single expert judge was found to display differential standards for these two groups of items. The judge in question was labeled 'rater one' and a plot of the mean ratings for this judge for items in both the focal and reference groups conditioned on the overall binary score is shown in Figure 1.

<insert Figure 1. about here>

Note that there appears to be reasonably uniform DPF in that this rater estimates that there is a greater probability of success for items in the area of Human Growth and Development *within subgroups of focal and reference items that were rated similarly by all judges.* That is, items in this content area were judged to be easier for the marginal group of examinees in instances where the overall item difficulties were rated similarly. Further, this DPF was unlikely due to chance (odds-ratio=0.051; $p<0.05$).

Example Two

The next data are ratings by committee members (14) of research proposals (25) submitted for critical review. The objects are the proposals while the raters are the committee members. A typical rater is presented (the rater is given number 'thirteen') where the question was whether committee

members were functioning differentially in their ratings of proposals from more technical areas (e.g., mathematics, biology, and chemistry; n=10) versus those from less technical areas (e.g., English, history, fine arts; n=15). There was concern about possible DPF since the technical papers required substantial specialized knowledge for a complete understanding. The plot is shown in Figure 2.

<insert Figure 2. about here>

An overall median rating (7.5 on a 1-10 scale) was used to create the dichotomous scores where a rating at or above the median is recorded as a '1' and a rating less than the median is recorded as a '0'. Five groups of similarly rated proposals (4-6 proposals in each group) were formed using these binary data. Figure 2. shows very similar lines for both the focal and reference groups. The odds-ratio (0.400) was not significantly different from one ($p>0.05$) and thus, we conclude that statistically significant evidence of DPF for this particular rater was not present. It was comforting to find that none of the raters in this example showed significant DPF between the more and less technical proposals.

Example Three

The final example data come from portions of an achievement test in mathematics administered to a group of 384 sixth graders participating in the evaluation of the *Lead Teacher Project* (Martin, 1989). The test consisted of 50 items on computation skills (mean item difficulty 0.513) and 55 items on comprehension and applications (mean item difficulty 0.592).

To identify possible DPF and construct plots similar to those in the previous examples, it was necessary to form item difficulty groups or clusters. Items from both sections of the test were grouped by difficulty level into seven clusters of similar difficulty. The number of items from both sections in each cluster ranged from 13 to 18; the mean item difficulties ranged from 0.166 in the first (most difficult) item group to 0.838 in the seventh (least difficult) item group.

Two persons were identified who had done better on the comprehension and application items than on the computation items. Person A had 56.4% of the comprehension and application items correct, but only 26.0% of the computation items correct. Figure 3. shows this person's

<insert Figure 3. about here>

performance on the similar item difficulty clusters of focal and reference groups of items. Note that person A shows significant (odds-ratio=0.203; $p<.05$) and uniform differential person functioning. That is, no matter what the item difficulty level, the performance of person A is better in the area of comprehension and application than in computation.

Compare this with the performance of person B who had 87.3% of the comprehension and application items correct and 74.0% of the computation items correct. Person B outperformed person A overall, but both individuals did less well on the more difficult computation items. Figure 4. would indicate that this seeming similarity is limited or even, perhaps, misleading.

<insert Figure 4. about here>

Person B did not perform in a differential manner across these two item types. Both the graph and the statistical test confirm that *within item groups of similar difficulty*, the score of person B was not significantly different (odds-ratio=0.698; $p>0.05$) between comprehension and application items and computation items. For person B, the performance difference between item types (or subscales on the test) is likely due to the fact that one type of item was easier than the other on this particular form of

the examination. For person A, the performance difference cannot be explained by item difficulty differences.

Discussion

The preceding examples have served to illustrate the potential utility of what we have termed *differential person functioning, DPF*, or the application of the principles of differential item functioning to rating data and transposed data from cognitive or diagnostic assessment. In two of our examples, the data sets were quite small and this does indicate the need for cautious interpretations. It is also true that some information was lost by artificially creating binary data. Alternative methods have recently been suggested for the detection of DIF using polytomously scored items (Millsap & Everson, 1993; Cohen, Kim, & Baker, 1993; Welch & Hoover, 1993).

Even if the notion of DPF proves unappealing, the lack of distinction between the concepts of *impact*, *differential functioning*, and *bias* on rating data and diagnostic testing is not informative. This mixing of person impact, DPF, and person bias would seem to contribute to the same sort of confusion that existed in the cognitive testing literature prior to the introduction of the more carefully crafted notion of DIF. A cautionary note is needed with respect to reasonably invariant (across both reference and focal groups) influences such as social desirability on the ratings of a judge or judges. It might well be justifiable to continue to refer to these effects simply as *rater bias* if the more precise DPF were used when appropriate.

The notions incorporated in DPF might prove especially helpful in the area of diagnostic testing. Certainly an individual displaying uniform DPF might require quite distinct assistance from one showing simple subscale differences as in our third example. A more complex situation might involve a non-uniform variety of DPF. For example, a person who was more successful on the easier computation items than on the comprehension and application items but who was less successful when items of both types become more difficult would be exhibiting that interaction of item difficulty and item type we might refer to as non-uniform DPF.

Finally, we note a few additional areas where these concepts might be applied. Differential person functioning could prove to be a useful construct in almost any area where rating scales are used. One example would be student evaluation of faculty. Researchers interested in studying the effects of fatigue on test-takers might find DPF useful in that the earlier items on an examination could be identified as the reference group of items and the later items as the focal group. If someone were suffering from fatigue during the testing, then you would expect to see both significant and uniform DPF. Specific rater characteristics such as differential functioning may be of ever increasing concern in the burgeoning area of performance assessments. In particular, many current assessments mix selected-response and constructed-response item formats. The possibility of DPF with respect to item format would seem an important area of concern for a variety of reasons ranging from the nature of the response pattern of the individual examinee all the way to the validity of the decisions arising from the assessment.

References

Angoff, W. H. (1984). Scales, norms, and equivalent scores. Princeton, NJ: Educational Testing Service.

Ansorge, C. J., & Scheer, J. K. (1988). International bias detected in judging gymnastic competition at the 1984 Olympic Games. Research Quarterly of Exercise and Sport, 59(2), 103-107.

Camilli, G., & Shepard, L. A. (1994). Methods for identifying biased test items. Thousand Oaks, CA: Sage Publications.

Cizek, G. J. (1996). Setting passing scores: An NCME instructional module. Educational Measurement: Issues and Practice, 15(4), 20-31.

Cohen, A. S., Kim, S. H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. Applied Psychological Measurement, 17(4), 335-350.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York, NY: Holt, Rinehart and Winston.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. Applied Measurement in Education, 2(3), 217-233.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland, & H. Wainer (Eds.), Differential item functioning (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer, & H. I. Braun (Eds.), Test validity (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Keller, B. G., & Bishop, R. C. (1985). Self-esteem as a source of raters' bias in peer evaluation. Psychological Reports, 56, 995-1000.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.

Martin, R. (1989). The lead teacher project: K-6 mathematics and science elementary teacher enhancement. Ohio University, Athens, OH. NSF grant number 91-47392.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. Applied Psychological Measurement, 17(4), 297-334.

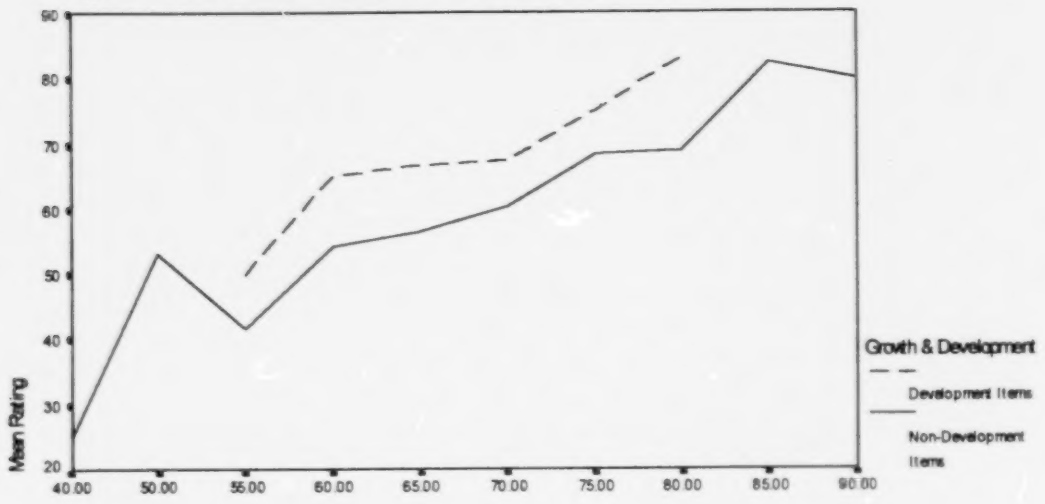
Nichols, D. P. (1994). The Mantel-Haenszel statistic for 2x2xK tables. Keywords: Tips and news for statistical software users, 54, 10-12.

Raju, N. S., Bode, R. K., & Larsen, V. S. (1989). An empirical assessment of the Mantel-Haenszel statistic for studying differential item performance. Applied Measurement in Education, 2(1), 1-13.

Uttaro, T., & Millsap, R. E. (1994). Factors influencing the Mantel-Haenszel procedure in the detection of differential item functioning. Applied Psychological Measurement, 18(1), 15-25.

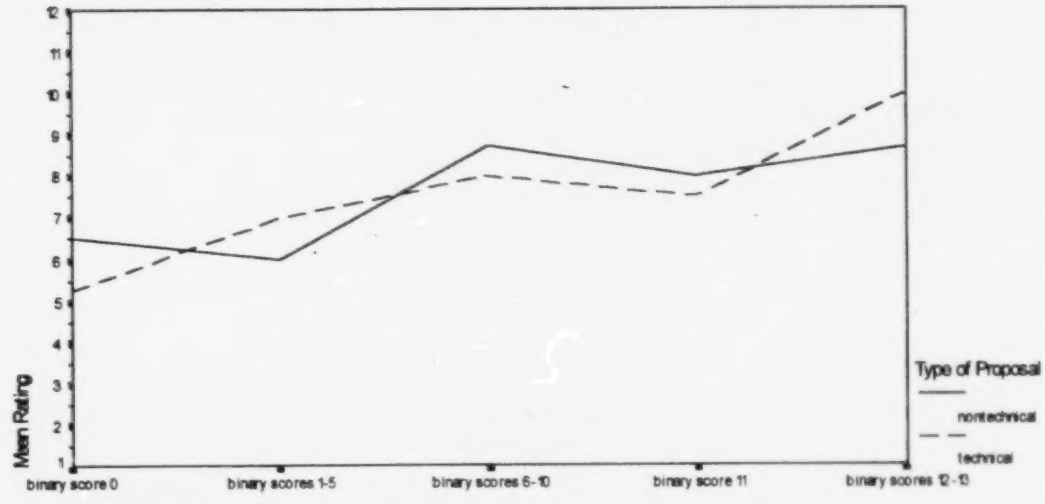
Welch, C., & Hoover, H. D. (1993). Procedures for extending item bias detection techniques to polytomously scored items. Applied Measurement in Education, 6(1), 1-19.

Figure 1. Focal-Reference Comparison* for Rater One



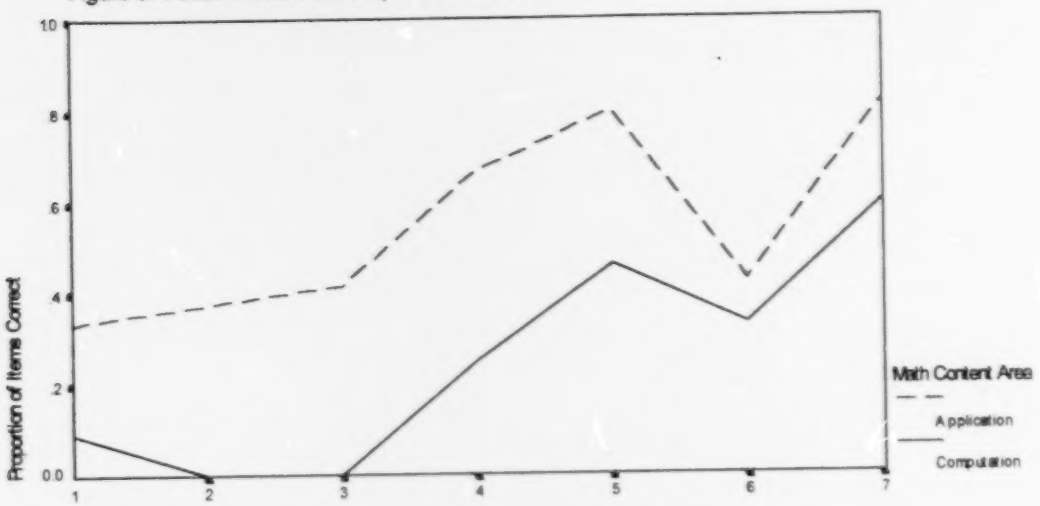
*MH chisquare is 6.065 with 1 df; p=0.014

Figure 2. Focal-Reference Comparison* for Rater Thirteen



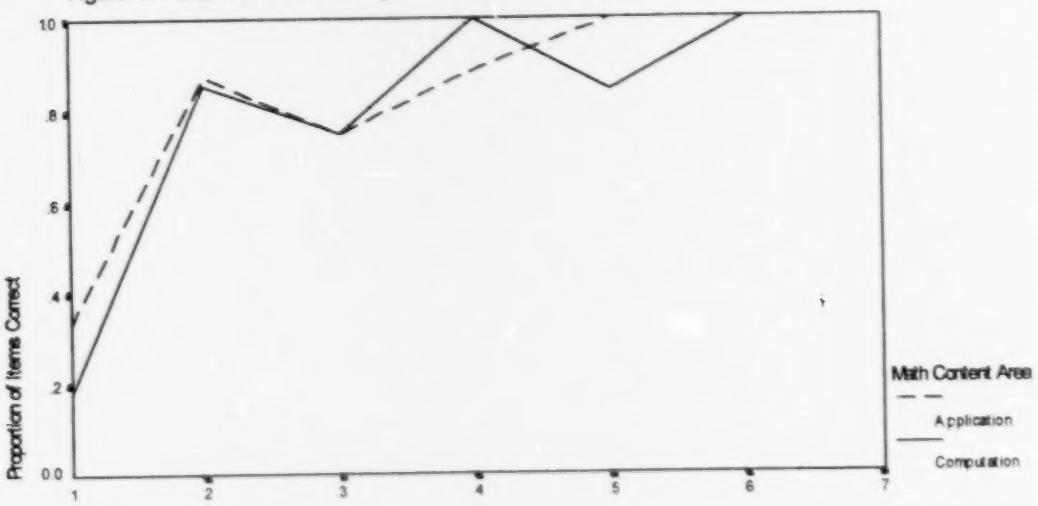
*MH chisquare is 0.098 with 1 df; p=0.755

Figure 3. Focal-Reference Comparison* for Person A



*MH chisquare is 7.821 with 1 df; p=0.005

Figure 4. Focal-Reference Comparison* for Person B



*MH chisquare is 0.014 with 1 df; p=0.905